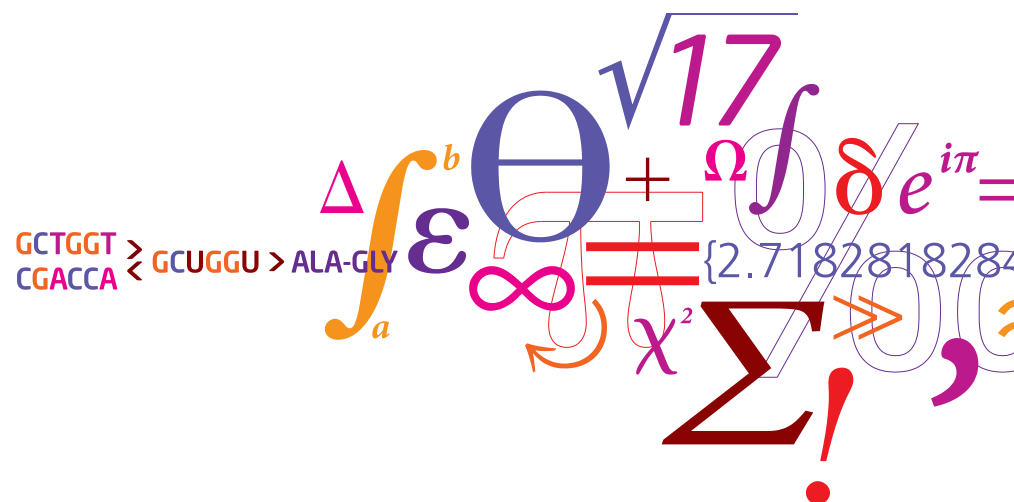


Variant Effect Prediction

Jose MG Izarzugaza, PhD

Integrative Systems Biology Group

Center Biological Sequence Analysis (CBS)



Understanding variants in *-omics* times

Traditionally

1 Mutation
=
1 Disease



Lots of hard work

Phenotype
Function
Mechanism

Now (High Throughput Sequencing, NGS)

X Mutations
In
Y Patients
And
Z Conditions




Prediction of
~~Unfeasibility /~~
Prioritization

Ensembl Variant Effect Predictor


http://www.ensembl.org/Homo_sapiens/Info/Index



[BLAST/BLAT](#) | [BioMart](#) | [Tools](#) | [Downloads](#) | [Help & Documentation](#) | [Blog](#) | [Mirrors](#)

Login · Register

Human (GRCh37) ▾







Human

Homo sapiens

Search Human...


e.g. [BRCA2](#) or [6:133017695-133161157](#) or [osteoarthritis](#)

Genome assembly: GRCh37 (GCA 000001405.11)

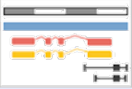
-  [More information and statistics](#)
-  [Download DNA sequence \(FASTA\)](#)
-  [Convert your data to GRCh37 coordinates](#)
-  [Display your data in Ensembl](#)

Other assemblies

- [NCBI36 \(Ensembl release 54\)](#)






[View karyotype](#)





[Example region](#)

Gene annotation

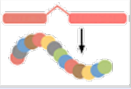
What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

-  [More about this genebuild](#)
-  [Download genes, cDNAs, ncRNA, proteins \(FASTA\)](#)
-  [Update your old Ensembl IDs](#)

 Additional manual annotation can be found in [Vega](#)





[Example gene](#)




[Example transcript](#)

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.



-  [More about comparative analysis](#)
-  [Download alignments \(EMF\)](#)




[Example gene tree](#)

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.





-  [More about the Ensembl regulatory build and microarray annotation](#)
-  [Download all regulatory features \(GFF\)](#)




[Example regulatory feature](#)


Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes.

-  [More about variation in Ensembl](#)
-  [Download all variants \(GVF\)](#)
-  [Variant Effect Predictor](#) 



[Example variant](#)



[Example phenotype](#)

Ensembl Variant Effect Predictor (I)



Variant Effect Predictor:

This tool takes a list of variant positions and alleles, and predicts the effects of each of these on overlapping transcripts and regulatory regions annotated in Ensembl. The tool accepts substitutions, insertions and deletions as input, see [data formats](#).



Upload is limited to 750 variants; lines after the limit will be ignored. Users with more than 750 variations can split files into smaller chunks, use the standalone [perl script](#) or the [variation API](#). See also [full documentation](#)

NB: Ensembl now by default uses Sequence Ontology terms to describe variation consequences. See [this page](#) for details

Input file

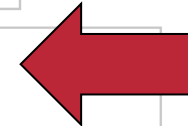
Species:

Human (Homo sapiens): GRCh37

Name for this data (optional):

Paste data:

```
1 881907 881906 -/C +
5 140532 140532 T/C +
```



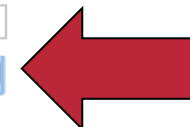
Upload file:

Choose File No file chosen

or provide file URL:

Input file format:

VCF



Ensembl Variant Effect Predictor (II)

Options

Transcript database to use:

- ☒ Ensembl transcripts
☐ RefSeq and other transcripts

Get regulatory region consequences (human and mouse only):



Type of consequences to display:

Sequence Ontology terms

Check for existing co-located variants:

Yes

Get 1000 Genomes global allele frequency for existing variants:



Return results for variants in coding regions only:



Show HGNC identifier for genes where available:



Show Ensembl protein identifiers where available:



Show HGVS identifiers for variants where available:

No

Missense SNP predictions (human only)

SIFT predictions:

Prediction and score

PolyPhen predictions:

Prediction and score

Frequency filtering of existing variants (human only)

Filter variants by frequency:



NB: Enabling frequency filtering may be slow for large datasets. The default options will filter out common variants found by the 1000 Genomes project.

Filter: Exclude variants with MAF greater than 0.01 in 1000 genomes (1KG) combined population

Next >

Ensembl Variant Effect Predictor (Results)

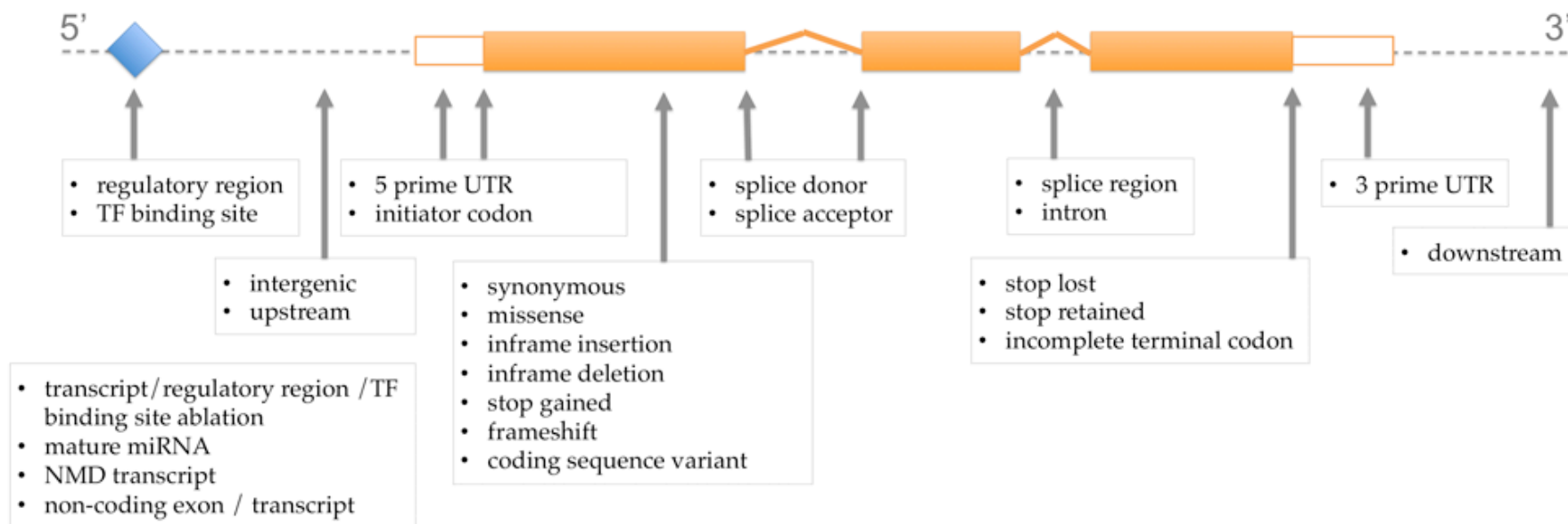


Variant Effect Predictor Results:

[Download text version](#)

Uploaded Variation	Location	Allele	Gene	Feature	Feature type	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
1_881907_-/C	1:881906-881907	C	ENSG00000187634	ENST00000466827	Transcript	downstream_gene_variant	-	-	-	-	-	-	DISTANCE=3724
5_140532_T/C	5:140532	C	ENSG00000249430	ENST00000512035	Transcript	downstream_gene_variant	-	-	-	-	-	rs12516846	DISTANCE=554; GMAF=C:0.1534
5_140532_T/C	5:140532	C	ENSG00000199540	ENST00000362670	Transcript	downstream_gene_variant	-	-	-	-	-	rs12516846	DISTANCE=3670; GMAF=C:0.1534
5_140532_T/C	5:140532	C	ENSG00000153404	ENST00000283426	Transcript	missense_variant	160	110	37	V/A	gTa/gCa	rs12516846	PolyPhen=benign(0); SIFT=tolerated(1); GMAF=C:0.1534
5_140532_T/C	5:140532	C	ENSG00000153404	ENST00000502646	Transcript	upstream_gene_variant	-	-	-	-	-	rs12516846	DISTANCE=149; GMAF=C:0.1534

Showing 11 to 15 of 15 entries

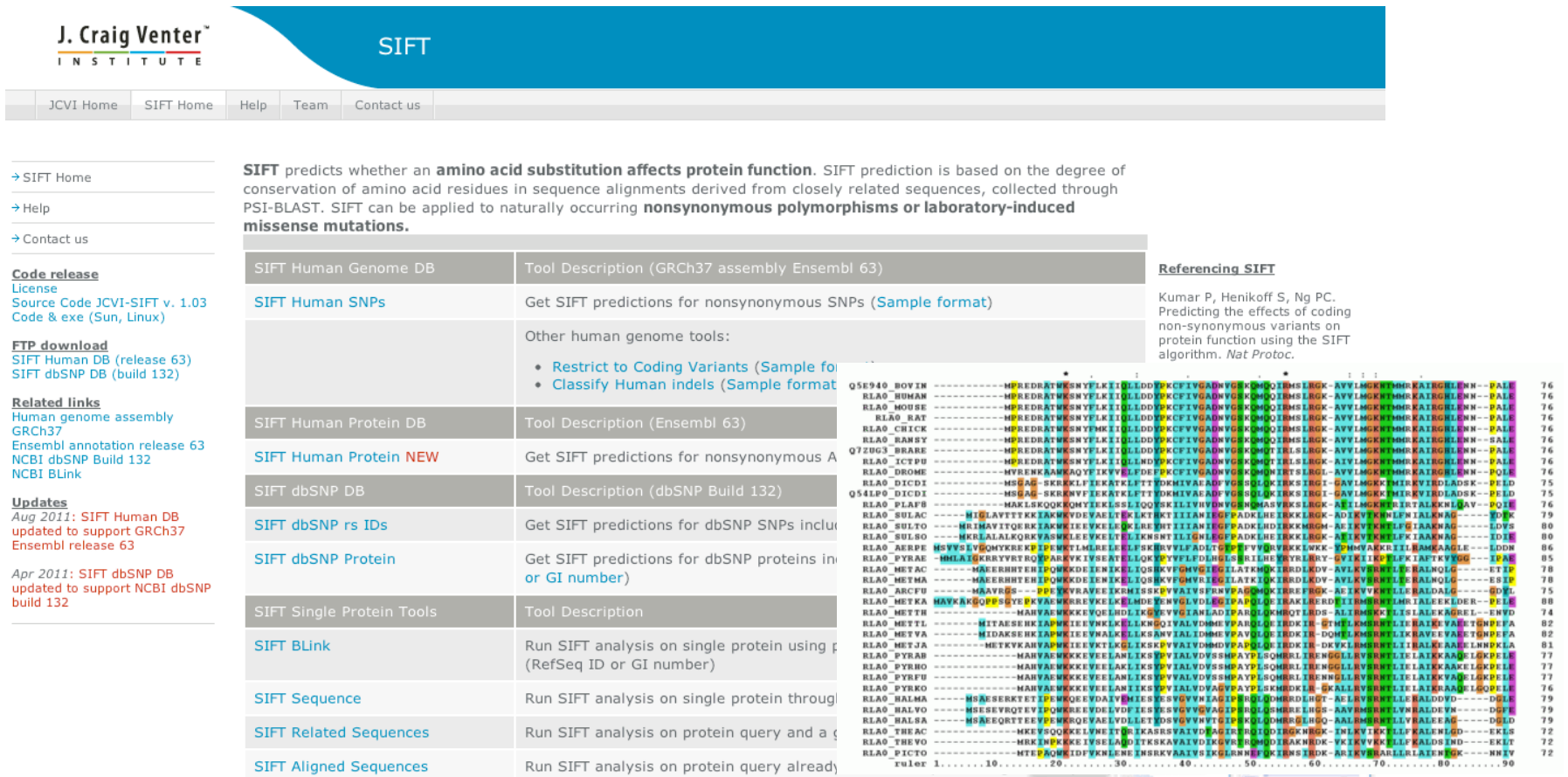


Predictors: SIFT

<http://sift.jcvi.org/>

Predicting Deleterious Amino Acid Substitutions

Pauline C. Ng and Steven Henikoff



- Based on the degree of conservation in a multiple sequence alignment (MSA)
- MSA generated from PSI-BLAST results (closely related sequences)
- Deleterious if $SIFT \leq 0.05$

Predictors: Polyphen-2



<http://genetics.bwh.harvard.edu/pph2/>

MACHINE LEARNING

- Naïve Bayes Classifier

A method and server for predicting damaging missense mutations

Ivan A. Adzhubei,^{1,7} Steffen Schmidt,^{2,7} Leonid Peshkin,^{3,7} Vasily E. Ramensky,⁴ Anna Gerasimova,⁵ Peer Bork,⁶ Alexey S. Kondrashov,⁵ and Shamil R. Sunyaev¹

SEQUENCE BASED FEATURES


- Importance of site: DISULFID, CROSSLNK, BINDING, ACT_SITE, LIPID, METAL, SITE, MOD_RES, CARBOHYD, NON_STD...
- Importance of region: TRANSMEM, INTRAMEM, COMPBIAS, REPEAT, COILED, SIGNAL, PROPEP...
- PSIC conservation score

STRUCTURE BASED FEATURES

- Likeness to destroy hydrophobic core, electrostatic interactions, interactions with ligands, or other important features of proteins

Predictors: PolyPhen-2





PolyPhen-2

prediction of functional effects of human nsSNPs

[Home](#) [About](#) [Help](#) [Downloads](#) [Batch query](#) [WHES.db](#)

PolyPhen-2 report for P15056 V600E

Query

Protein Acc	Position	AA ₁	AA ₂	Description
-------------	----------	-----------------	-----------------	-------------

P15056	600	V	E	Canonical; RecName: Full=Serine/threonine-protein kinase B-raf; EC=2.7.11.1; AltName: Full=Proto-oncogene B-Raf; AltName: Full=p94; AltName: Full=v-Raf murine sarcoma viral oncogene homolog B1; Length: 766
------------------------	-----	---	---	---

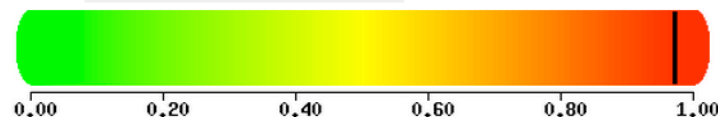
Results

☒ Prediction/Confidence

PolyPhen-2 v2.2.2r398

HumDiv

This mutation is predicted to be **PROBABLY DAMAGING** with a score of **0.971** (sensitivity: **0.77**; specificity: **0.96**)



☒ HumVar

Details

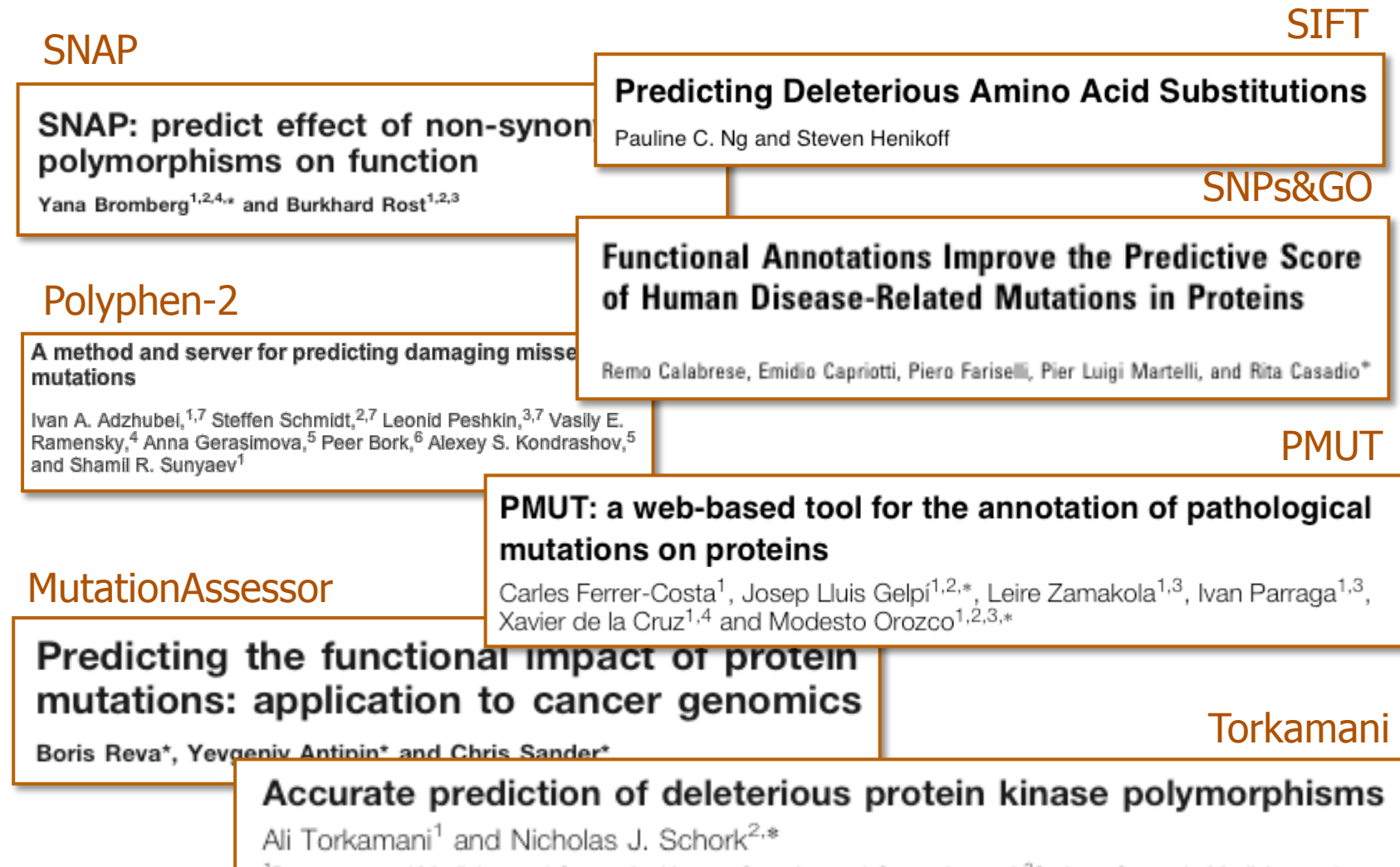
☒ Multiple sequence alignment

UniProtKB/UniRef100 Release 2011_12 (14-Dec-2011)

QUERY	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp G1P9K1#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp B7ZRT9#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp Q0D2E4#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp G1NKK9#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp Q68FI8#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp Q4F9K6#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp Q643Z8#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp Q767H5#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp G3Q6E4#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp G3Q6E7#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp UPI00016E35C7#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp G3Q6E5#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp UPI00017B47FE#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp UPI00017B47FF#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp B3DFX5#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp Q1LYG2#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV
sp UPI00017B4800#1	KS--IIHRDLKSNNIFLHED-L---TVKIGDFGLATV	KSRWS---GSHQFEQ-----LSGSILWMAPEV

Shown are 75 amino acids surrounding the mutation position (marked with a black box). An interactive version of the complete alignment is [also available](#).

Automatic methods to predict the pathogenicity of mutations



Some of the (many) methods implemented during the last decade

Predictors: SNPs&GO

<http://snps-and-go.biocomp.unibo.it/>

MACHINE LEARNING

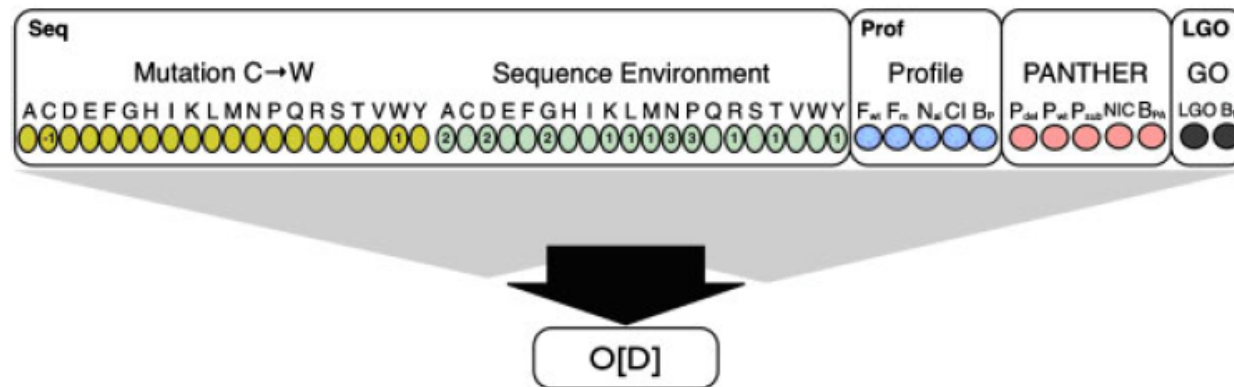
- Support Vector Machine

FEATURES

- Wild type and mutant amino acids
- Sequence environment (20 slots, 9-x-9)
- Profile information: Conservation in MSA (freq. of Aa, Depth, Conservation index)
- PANTHER prediction (prob. Being disease, prob. wt Aa and mut Aa in pos)
- Gene Ontology Log-odds ratio

Functional Annotations Improve the Predictive Score of Human Disease-Related Mutations in Proteins

Remo Calabrese, Emidio Capriotti, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio*



Predictors: SNPs&GO



<http://snps-and-go.biocomp.unibo.it/>

SNPs&GO

Predicting Human Disease-related Mutations in Proteins with Functional Annotations

[Home](#)
[Information](#)
[Help](#)
[About us](#)
[Reference](#)
[Useful Links](#)

Welcome

SNPs&GO is a server for the prediction of single point protein mutations likely to be involved in the insurgence of diseases in humans.

UNIPROT Accession Number

Mutation Position

Wild-type residue

Substituting residue

For further information and bug report please contact: gigi@biocomp.unibo.it

Predictors: SNPs&GO



<http://snps-and-go.biocomp.unibo.it/>

SNPs&GO

Predicting Human Disease-related Mutations in Proteins with Functional Annotations

[Home](#)

[Information](#)

[Help](#)

[About us](#)

[Reference](#)

```

                                SNPs&GO Prediction
*****
**                                                                    **
**                                RESULTS                                **
**                                                                    **
*****
SEQ File: P15056
Position  WT  NEW      Effect  RI
        600   V   E      Disease  7

BP: GO:0006468 BP: GO:0006916 BP: GO:0007242 BP: GO:0009887
MF: GO:0000166 MF: GO:0004674 MF: GO:0005057 MF: GO:0005515 MF: GO:0005524 MF: GO:0008270 MF: GO:0016740 MF: GO:0019992
CC: GO:0005737 CC: GO:0005886

WT: Residue in Wild-Type Protein
NEW: Mutated Residue
RI: Reliability Index
Effect:
      Neutral: Neutral Polymorphism
      Disease: Disease-related Polymorphism
BP: Biological Process GO term
MF: Molecular Function GO term
CC: Cellular Component GO term

*****
**                                                                    **
**                                http://snps-and-go.biocomp.unibo.it/snps-and-go/                                **
**                                                                    **
*****
```

ePIPE: Prediction of protein features

